IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 February 2024, accepted 12 March 2024, date of publication 19 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3378105

RESEARCH ARTICLE

VKP-P3D: Real-Time Monocular Pseudo 3D Object Detection Based on Visible Key Points and Camera Geometry

CHANGLIANG SUN, HONGLI LIU, (Member, IEEE), WEICHU XIAO, BO SHI, AND YUAN QIU
College of Electrical and Information Engineering, Hunan University, Changsha 410082, China

Corresponding author: Changliang Sun (changliang.sun@hnu.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 62173133.

ABSTRACT Three-dimensional object detection has been substantially improved with the use of expensive LiDAR and stereo vision systems in intelligent driving. The less-expensive and more scalable solution of monocular 3D object detection, however, remains a key challenge. This study primarily explores real-time pseudo 3D object detection with monocular vision and designs a single-shot RPN model, VKP-P3D, which relies purely on visual feature extraction. Through a multiscale feature fusion and an attention mechanism module, this model obtains high-dimensional feature representations during the feature extraction phase. In the detection head of the VKP-P3D model, the pseudo 3D object detection is obtained by regressing 2D bounding box and the visible key points within the image coordinate of the 3D box from the camera's perspective. Finally, assuming flat ground and considering geometric parameters of the camera, the object's 3D information can be extracted. To verify the effectiveness of the proposed algorithm, we constructed two pseudo 3D object detection datasets based on visible key points and compared with current state-of-the-art real-time object detector. Results showed that the proposed model has high detection accuracy and speed.

INDEX TERMS Monocular vision, pseudo-3D object detection, visual key point, single-shot RPN network.

I. INTRODUCTION

In the field of intelligent driving perception, the adoption of cameras has gained popularity due to their affordability and rich information. In the early stage, cameras were used to provide drivers with blind spot imagery. At present, these cameras serve as crucial sensors for detecting obstacles, traffic signals, and road markings within the advanced intelligent driving perception system.

In the monocular visual perception domain, the research on 2D object detection is relatively mature. The main approach is to detect the object's 2D bounding boxes of vehicles, pedestrians, and riders and predict the motion path for the decision module of intelligent driving systems. Progress in perception algorithms has come a long way from initial stages of pedestrian detection [1] and traffic sign recognition [2] based on feature extraction engineering and machine learning, to the use of deep neural network to achieve the multiobjective and high-precision task in complex scenes [3], [4], [5], [6], [7], [9], [10]. This evolution has drastically improved the functionality and performance of the intelligent driving perception system. However, the utility of 2D bounding box information has been unable to meet the demand for advanced intelligent driving system.

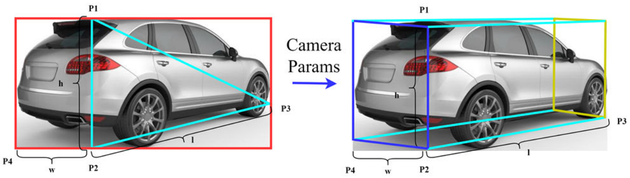

FIGURE 1. This paper construct a single-shot RPN network to recognize the object's pseudo 3D information with monocular image, and then converted to 3D bounding box based on the camera parameters and ground flat assumption.

Extracting object's 3D information from 2D image based on visual features is a popular research domain in recent

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

years. However, in comparison with the 3D point cloud data from LiDAR [18], [19], [20], [21], [22], [23] or stereo vision [24], [25], monocular images can only provide 2D information. The extraction of 3D features from images is a major challenge for visual-based algorithms, especially in the field of monocular 3D object detection.

In the application of intelligent driving perception, methods such as monocular depth regression designing feature encoding and decoding structures to regress the depth information of the perception field of camera [15], [16], [17]. These methods drastically reduce the performance in complex driving scenarios (e.g., backlight, low light, and occlusion); moreover, the construction of a large-scale depth dataset requires LiDAR point cloud data for annotation, which is costly [12], [13]. Other methods [27], [32] combine 2D object detection and monocular depth information to calculate the object's 3D bounding box in the camera coordinate system. These methods require high-precision 3D object annotation with LiDAR point cloud to construct a dataset, and the object's 3D information of the monocular image is regressed while training the model.

This paper mainly studies the monocular pseudo 3D object detection algorithm with pure visual features in intelligent driving system. The proposed model does not rely on LiDAR point cloud data for dataset construction, model training, and real-time inference processing. It combines the present cutting-edge object detection technology and designs a real-time pseudo 3D object detection model based on visible key points (see Fig. 1). We optimize the feature extraction network model to improve the detection accuracy. Fig. 3 illustrates an overview of our proposed framework. To validate the proposed algorithm, we construct two pseudo 3D object detection datasets based on visible key points for experimental verification, namely, StellaIR-P3D and KITTI-P3D. Experimental results show that the proposed method has fast detection speed and high accuracy. The main innovations of this paper are as follows.

1) This study designs a monocular pseudo 3D object detection framework involving a single-shot RPN structure for an intelligent driving system. This framework maps the object's 3D bounding box vertices in the camera coordinate to visible key points in the image coordinate system. By regressing the 2D box and 3D visible key points of an object, the end-to-end, real-time pseudo 3D object detection model is established.
2) The StellaIR-P3D and KITTI-P3D datasets are constructed. Annotations for these datasets do not require the use of LiDAR point cloud information and provide richer pseudo 3D information compared with 2D bounding box datasets.
3) By combining multiscale feature fusion, a self-learning attention structure is designed and implemented to improve the feature extraction performance of the backbone network for complex scenes.
4) Assuming that the ground is flat and applying the principles of monocular imaging geometry, we convert the visible key points from the pseudo 3D box in image coordinates into a 3D pose in camera coordinates.

II. RELATED WORKS

The sensors used in the intelligent driving perception system for 3D object detection mainly include LiDARs and cameras. The point cloud data of LiDAR makes it easy to obtain the 3D information of objects, but the resolution is typically low and the cost is considerably higher than other sensors. Moreover, point cloud data are easily affected by adverse conditions such as rain, fog, or dusty environments. Conversely, cameras, while providing richer resolution, present challenges in perceiving 3D information.

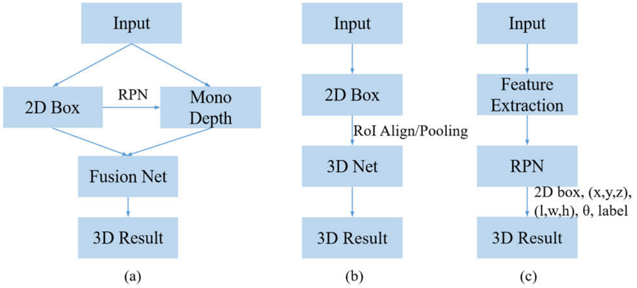

FIGURE 2. Current mainstream of monocular 3D object detection architecture. References [27], [29], [32], and [33] are comprised of a monocular depth network to obtain point cloud in (a). [30], [31], and [34] adopt two stage structure to conduct 2D bounding box detection followed by regression of 3D information in (b). [26] and [28] design an end-to-end network and obtain 3D object detection based on monocular visual geometric parameters and ground information and it is also the detection architecture that this paper focuses on.

A. 3D OBJECT DETECTION BASED ON MULTISENSOR FUSION

In recent research of intelligent driving system of 3D object detection, fusion solution [21] combining multiple heterogeneous sensors, such as cameras [19], LiDARs [18], radars, has been profoundly studied and applied, these approaches capitalize on the unique advantages of each sensor, thereby improving recognition accuracy. Reference [20] initially calculates the transformation relation between the point cloud and image pixels through coordinate calibration. Then, it uses image object detection algorithms to combine point cloud data to recognize object 3D information. However, this solution is usually used for autonomous driving perception systems, and its cost is usually high. Reference [22] proposes a fast yet effective backbone, termed VirConvNet, based on a new operator VirConv (Virtual Sparse Convolution), for virtual-point-based 3D object detection. This structure alleviates the computation problem by discarding large amounts of nearby redundant voxels and tackles the noise problem by encoding voxel features. For the problem of object occlusion under various practical conditions, [23] introduce a point reconstruction network module designed to recover the missing 3-D spatial structures of foreground points.

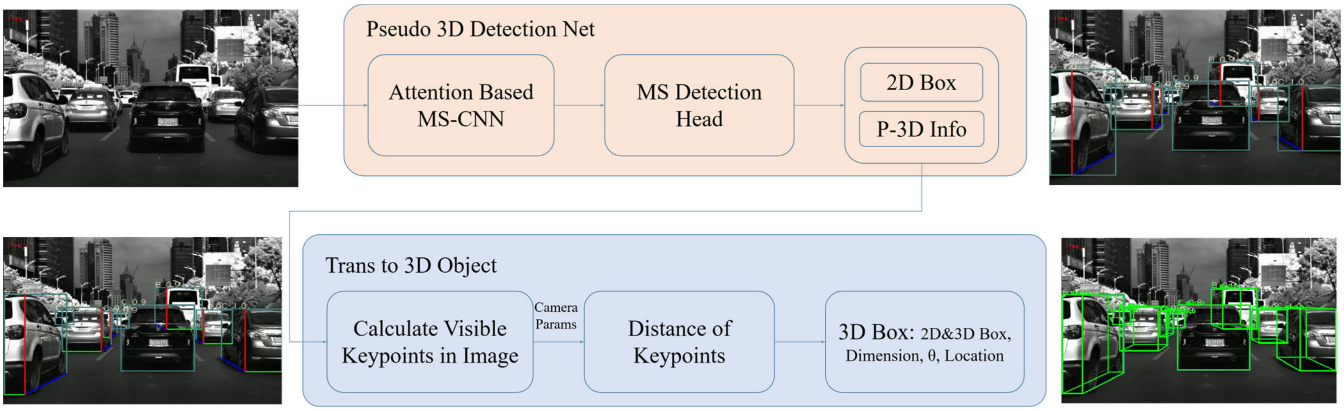

FIGURE 3. Overview of the proposed monocular 3D object detection framework. 2D bounding box and pseudo-3D information is firstly recognized by a single shot network, then calculates the visible vertices of the object's 3D box, and resolves the 3D position in camera coordinate based on geometric imaging principles.

Another pivotal technique for achieving 3D object detection uses stereo vision [24], [25]. This technology operates by positioning two identical cameras side by side, ensuring that the optical axes are parallel. The synchronous images from the two cameras then undergo pixel matching, and the distance is calculated based on the geometric relationship between the two cameras' perspective difference. This technology can obtain 3D point cloud information with high accuracy of the perception range under appropriate conditions. However, the disadvantage is that it requires high accuracy hardware and calibration precision. It is also difficult to recover the depth information for some special scenarios such as low illumination, glare.

B. 3D OBJECT DETECTION VIA MONOCULAR VISION

Due to their high performance-to-cost ratio and scalability, monocular vision-based 3D object detection algorithms have found broad utilization in the sphere of intelligent driving. The prevailing technical approach relies on regressing the 3D information of objects based on 2D bounding boxes [26], [27], [28], [29], [30], [31], [32], [33], [34]. Two main solutions can be used to achieve this. One involves designing a monocular depth regression network to obtain the pseudo LiDAR point cloud from the image. For instance, RoI-10D [32] uses three network models for monocular 3D object recognition. Of these, one network uses ResNet-FPN for 2D bounding box object detection, and another uses monocular depth estimation with off-the-shelf depth prediction networks. The detection network generates a box and uses RoI align to concatenate features with the depth result. A new feature map then transitions from 2D to 3D using RoI lifting net. In a similar vein, AM3D [27] performs monocular 3D object detection using three networks—one for monocular depth estimation, another for 2D bounding box object detection, and a final one for regressing 3D information. The detection network generates a 2D box and integrates it with the depth map using RoI align method for point cloud fusion. Then, a detection network, Det-Net, is designed to recognize the object's 3D pose (e.g., position, dimension, and global angle) based on multidimensional features composed of point cloud and RGB pixels. Notably, the aforementioned algorithms require the regression of a monocular depth map, resulting in poor real-time performance.

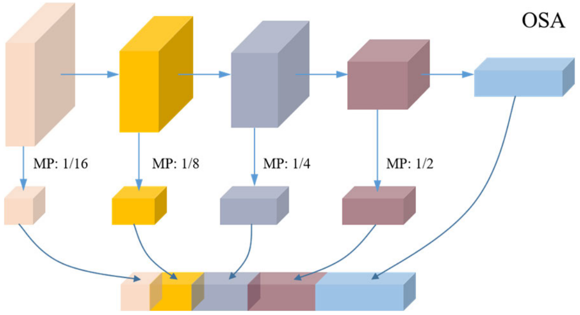

FIGURE 4. OSA(One-Shot Aggregation) structure integrates features of different hierarchies and scales to enrich the granularity of features.

Another solution involves directly regressing a 3D bounding box in one object detection framework. These frameworks calculate the 3D information of the object with premise of a flat ground (ignoring roll angle) and camera geometric characteristics. Reference [33] based its calculations on the assumption of the ground being flat, thus assuming that for all points on the ground, the depth is known. However, for objects elevated off the ground, depth cannot be directly calculated. Instead, using statistics from the training dataset, it can obtain the mean variance of the depth and observation angle between the object's central point and ground points for each candidate box, to be used as prior information. On this basis, a GAC module is proposed for 3D object data regression in [33]. In [26], the MS-CNN [10] object detection model was used for 2D object detection, and the feature maps of the 2D object were extracted. Three fully connected neural networks were used to regress the length, width, height, angle, and confidence. Reference [34] transformed 3D object detection into a 2D object detection module

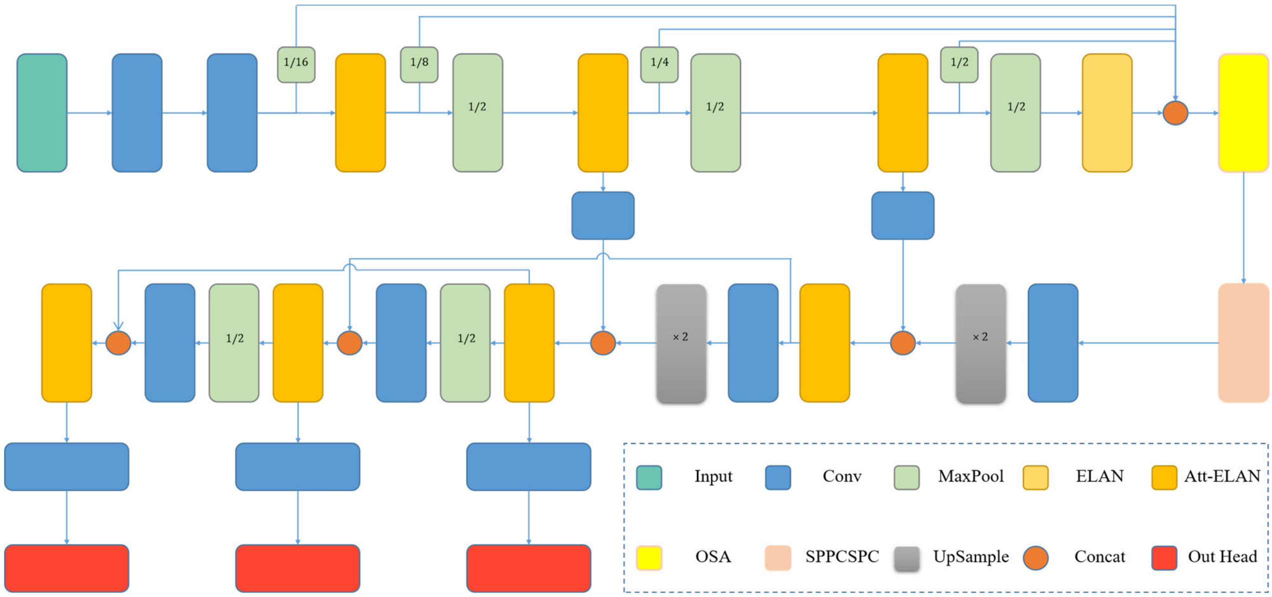

FIGURE 5. The proposed monocular pseudo 3D object detector refers to the current state-of-the-art real-time 2D object detector. The multi-scale feature fusion structure is adopted in backbone network, and the detection head has three scale outputs to detect different size objects. The outputs include 2D bounding box and pseudo 3D information.

and 3D box vertex recognition module by regressing box vertices. This algorithm adds a branch on the original object detection module to predict the coordinate of each vertex. A comparison between several major structure of monocular 3D detector is presented in Fig. 2.

However, these types of method rely on LiDAR annotation in their training datasets to obtain distance, length, width, and other 3D object data. All these algorithms are designed and implemented based on high-precision 3D information datasets annotated with LiDAR point cloud. For example, in the KITTI monocular 3D object detection dataset, cameras and LiDARs provide the 2D bounding box coordinates for each object from the image. The object's distance, size, global angle, and other 3D information are mainly obtained through LiDAR point clouds. Then, it constructs a monocular 3D object detection dataset based on the camera coordinate with the 3D information of the objects.

In intelligent driving scenarios, the complexity of the perceptual environment necessitates large datasets to train robust detection algorithms. The construction of a 3D object detection dataset based on cameras and LiDAR fusion annotation is highly expensive, and problems such as calibration accuracy, data synchronization, and noise reduction must be solved. Therefore, constructing a pseudo 3D object detection dataset using monocular visual annotation provides richer perceptual information compared with monocular 2D bounding box object detection datasets. Furthermore, in comparison with monocular 3D object detection datasets based on point cloud-assisted annotation, the cost is lower, making it a valuable asset for intelligent driving perception technology.

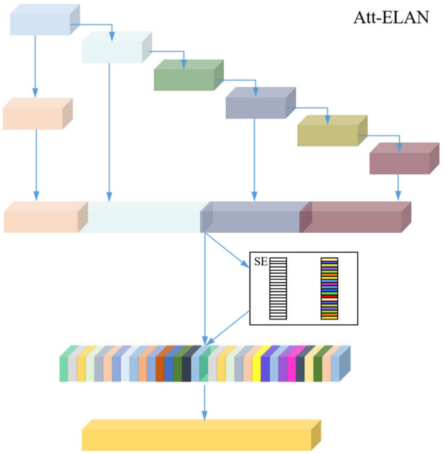

FIGURE 6. Proposed Att-ELAN structure: Adopting SE block for multi-feature channels to enhance the important characteristics.

III. METHODOLOGY

This paper proposes a model based on a single-shot RPN network called VKP-P3D, which uses purely visual features for monocular 3D object detection tasks. Multiscale fusion and self-attention features are initially extracted at the backbone network. The prediction module regresses the 2D object bounding box and visible key points of the 3D box in the image coordinate system. Subsequently, on the basis of the camera's geometric principle and flat ground assumption, the object's visible key points of the 3D box are converted into the camera coordinate to obtain the object's 3D information, including distance, dimension, and global orientation angle. The main parts of proposed approach are shown in Fig. 3.

A. ATTENTION-BASED MULTISCALE FEATURE EXTRACTION NETWORK

In comparison with object classification, detection tasks need more diverse scales to recognize objects. Retaining information from each network layer during model training is essential because each layer of the network has different receptive fields.

In recent research on feature extraction networks, many models have improved the expression ability by aggregating features at different levels [35], [36], [37]. For example, DenseNet [39] uses dense connections to aggregate intermediate features with different receptive field sizes, which enables feature reuse at different levels but reduces the speed and efficiency of model operation. VoVNet [40] aggregates the features of all previous layers only at the last layer of the module through the one-shot aggregation (OSA) structure, which improves the computing efficiency while obtaining different granularity features. In YOLOv7 [7] object detection backbone network, multiple E-ELAN structures [38] are used to fuse features between different levels, which can enhance the features learned from various feature maps and improve parameter utilization and computational efficiency. In this paper, our backbone network incorporates feature extraction structures from the aforementioned algorithms.

The self-attention mechanism is an important research field in the feature extraction module. In the context of intelligent driving applications, dataset images can exhibit significant differences due to environmental factors such as lighting, weather conditions, and scenes. Therefore, feature extraction modules must possess strong generalization ability. SENet [41] introduces a novel architectural unit, namely, the ''Squeeze and Excitation'' (SE) block, to enhance the performance of neural networks by selectively emphasizing informative features. It enhances useful channels with global features while attenuating invalid channels to improve the representational ability of neural networks.

In terms of feature extraction in the backbone network, this paper integrates feature aggregation and key feature self-learning mechanisms (see Fig. 5). The proposed network adopts feature fusion structure at different levels and scales to obtain feature expressions at multiple granularity spaces. Then, based on the self-learning attention structure, the importance of each channel is calculated and screened to evaluate the channel contribution. The SE block automatically allocates low weights to weak feature channels and improves the feature expression ability in complex scenes with reduced computational complexity. Research results show that these structures can effectively improve recognition accuracy.

1) OSA

To aggregate features at different resolution maps, this paper adds the OSA module to the backbone feature extraction network. The five shallow features extracted from the front-end levels are downsampled into the same resolution using MaxPool operation and then connected through the Concat layer to fuse different granularity features. The OSA structure is shown in Fig. 4. By adding this structure to the model, the GMac value increases by 0.8%, thereby improving the mean average precision (mAP) of the model by 0.1%.

2) ATT-ELAN STRUCTURE

A multifeature fusion structure, proposed in [38], continuously improves the learning ability of a network without damaging the original gradient path. In this paper, this feature extraction structure is fused with the self-attention mechanism. The SE block is added to the ELAN feature. The weight of each channel of the fused features is automatically calculated, and the dominant channels are enhanced to improve the expression ability of feature extraction models. The Att-ELAN structure is shown in Fig. 6. This structure increases the GMac by 2.2%, and the mAP of the detection model is improved by 1.7%.

B. PSEUDO 3D PREDICTION MODULE BASED ON VISUAL KEY POINTS

In intelligent driving scenarios, various studies have endeavored to obtain an object's 3D information via a monocular camera, particularly in real-time 3D object detection research. The monocular 3D object detection task examined in this paper focuses on pure visual perception, without the assistance of LiDAR data annotation. The object's 3D information is obtained through the visible key points of the 3D bounding box in the image coordinate system. A single-shot RPN structure, designed with fully convolutional neural network, is incorporated. The main prediction content of the VKP-P3D detection head includes two parts: the 2D bounding box and the 3D box visible key points in the image coordinate system.

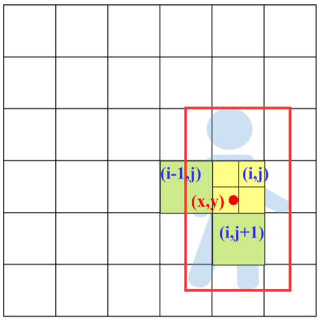

FIGURE 7. The center point (x, y) of the person object's 2D bounding box is located at the left bottom corner of the yellow grid (i, j), the regions corresponding to the yellow cell (i, j) and green cells (i-1, j), (i, j+1) will be assigned as the positive ground truth. The number of positive samples in the model output is increased.

For 2D object bounding box detection, tiny object detection is always a hot spot in intelligent driving perception system. Reference [11] propose a few-shot image object detection with confidence-iou collaborative proposal filtration and tiny object constraint loss (FsCIT). This strategy is effective and can improve the performance of tiny object detection. In this paper, we combine the existing mature anchor-based algorithm ideas, design a model with the fully convolutional neural network, and output three different scale feature maps. The feature maps of different scales correspond to the objects of differing sizes. The cell unit of each feature map is used to predict the existence probability, category, and 2D bounding box of an object.

In terms of pseudo 3D information regression, this paper extends the 2D bounding box detection and transforms the 3D object detection into visible key points regression tasks. It mainly calculates four aspect information: vehicle front orientation (VFO), vehicle front or rear proportion (VFRP), vehicle traveling orientation (VTO), and VTO proportion (VTOP). These aspects allow us to obtain the visible key points of objects. On the basis of the principle of monocular imaging and assumption of flat ground, the 3D object distance, dimension, and global orientation angle are calculated based on the visible key points of the 3D box to implement the transformation from pseudo 3D visible key points to 3D bounding box.

1) 2D BOUNDING BOX DETECTION

The approach 2D object detection in this paper relies on an anchor-based mechanism, which mainly regresses the object's category, confidence, and the bounding box. For each cell of detector feature maps, three anchors with different aspect ratios and sizes are used to predict objects of different sizes. We refer to YOLOv7 [7] for the loss function L_{2d} for training the 2D bounding box.

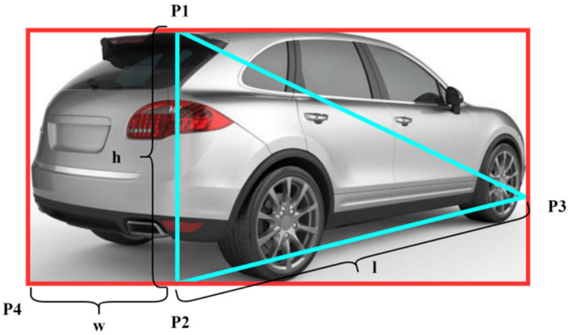

FIGURE 8. P1, P2, and P3 are three annotated vertices. Based on the ground flat assumption, (P1, P2) is perpendicular to the ground, corresponding to the vehicle's height; (P2, P3) is on the ground, corresponding to the vehicle's length; (P2, P4) is on the ground and approximately represents the vehicle's width.

In object detection task, the number of positive samples is considerably less than that of the negative samples. Only the loss from positive samples is calculated during the training process. To improve the recall precision of the detector, label assignment strategies, which aim to assign each cell of training sample a positive or negative loss, plays an important role in object detection [45]. In terms of object candidate regions, this paper adopts a four-neighborhood multiple label assignment strategy. For each grid of feature maps, where the center of the object's ground truth is located, the adjacent grid also has the ability to detect this object. Therefore, the three nearest grids where the object center is located is assign as positive samples to calculate the training loss (see Fig. 7). This strategy can effectively increase the number of positive samples and improve the object detection ability of the proposed model.

In terms of object classification, the soft label allocation strategy is considered an effective means to combat overfitting. Label smoothing was first proposed for classification problems in [43]. Previously, the labels of object classification task were described using one hot encoding. During the training process, if it is considered to belong to the nth classification, then the position output is 1, and the others' positions are 0. This allocation strategy is prone to overfitting issues. References [43] and [44] generated soft labels through the label smoothing regulation strategy to improve model adaptability.

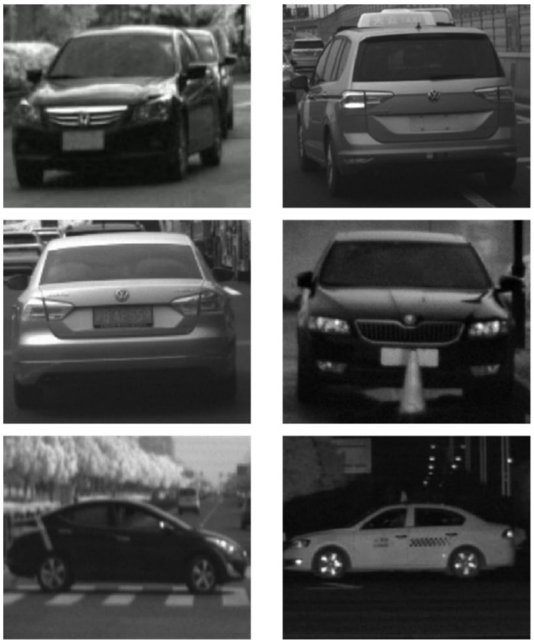

FIGURE 9. The top two images depict typical vehicle postures. In left image, VFO is backward relative to the ego vehicle and VTO is left relative to the image perspective. In right image, VFO is forward and VTO is left. For the two middle images, assign the categories of VTO (left, middle or right) with the label smoothing strategy is beneficial for preventing overfitting. For the bottom two images, assigning the categories of VFO (forward or backward) using soft labeling method is also necessary.

In this paper, we initially train the model through hard label allocation method to obtain the confusion matrix of object categories. Then, according to the confusion relationship of different categories, we assign soft labels to fine tune the model and improve the generalization ability of the detector. This strategy increases the accuracy by 0.2% without adding any extra burden of this model.

2) PSEUDO 3D INFORMATION REGRESSION

The main innovation of this paper is the pseudo 3D information regression based on visible key points of the 3D bounding

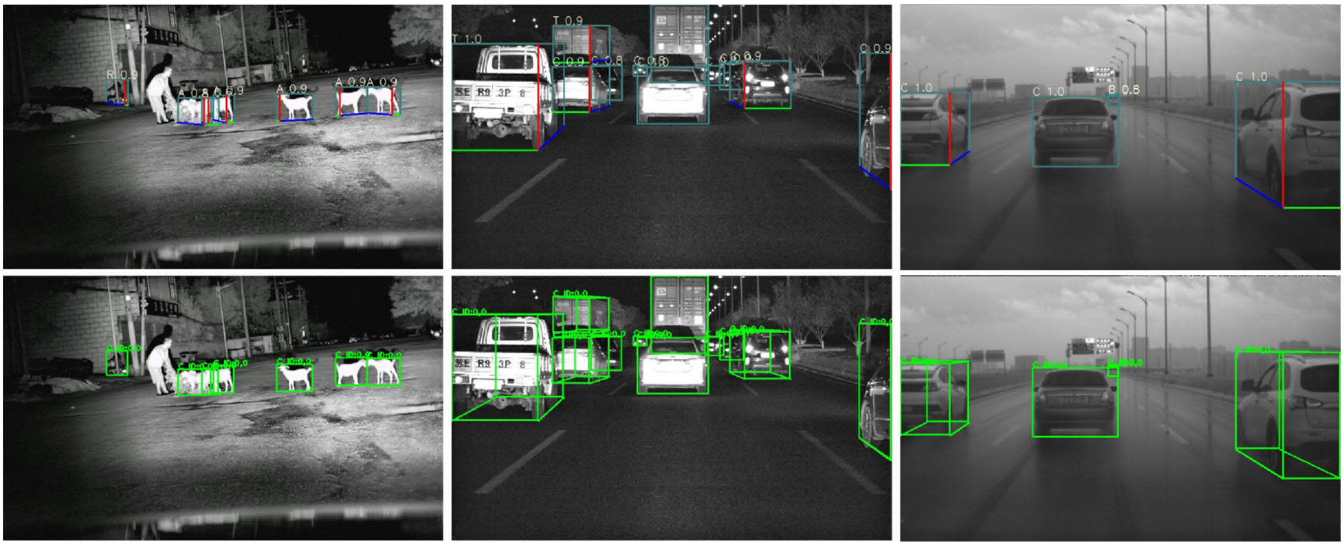

FIGURE 10. We visualize qualitative examples of our proposed method for multi-class 3D object detection in StellaIR-P3D dataset collected by IR sensor. The top two rows visualize the pseudo 3D object detection results in different scenarios include daytime, rainy, night with low light. The corresponding 3D bounding boxes results are demonstrated in the bottom two rows.

box. In comparison with existing monocular 3D object detection model from the aforementioned algorithms, this paper transforms the 3D object detection task into regressing the visible key points of the object's 3D bounding box in the image coordinate system.

As shown in Fig. 8, from the current image perspective, the visible key points of the vehicle's 3D bounding box are approximately p1, p2, p3, and p4. The segments of these vertices correspond to the length, width, and height of the object's 3D bounding box. Therefore, we simply predict the coordinate of these visible key points in the image coordinate using the VKP-P3D network and transform them with camera geometric parameters and flat ground assumption to calculate the vehicle's 3D pose.

In terms of key point detection, numerous studies have been conducted on facial key point detection [46] and human pose estimation [47], [48]. However, in the application of intelligent driving vehicle detection, the recognition accuracy through directly regressing eight vertices of a 3D bounding box is unsatisfactory [34]. This paper transforms the problem of visible key points detection of a vehicle's 3D bounding boxes into regressing the VFO, VFRP, VTO, and VTOP. Subsequently, it solves the coordinates of the visible key points in the image.

a: VFO

We categorize the VFO in image perspective into two categories: toward the ego car and away from the ego car. In most situations, the categories of VFO have two properties during labeling: forward and backward. However, in the experimental dataset, when the vehicle is in lateral pose in the image, different annotators may annotate inconsistent label for the same object's pose, as shown in Fig. 9. To optimize the allocation of VFO classification labels, a label smoothing strategy is used, which weakens the certainty of difficult-to-distinguish VFO ground truth categories and reduces overfitting during model training. During the training process, an online soft label allocation strategy is used for VFO classification when VFRP is close to zero, instead of allocating labels with one-shot encoding. The formula for the VFO classification soft label is defined as follows:

$$p'(k|x_i) = (1-\alpha)\cdot p(k|x_i) + \alpha/2, \tag{1}$$

$$\alpha = \begin{cases} 0 & if\ w_p > 0.1 \\ max(1-10*w_p, 0) & otherwise, \end{cases} \tag{2}$$

where $\mathrm{x_i}$ denotes a sample with $\mathrm{y_i}$ label; $\mathrm{p(k|x_i)}$ is the one hot label assigned with $\mathrm{p\,(k = y_i \mid x_i)} = 1, \mathrm{p(k \neq y_i|x_i)} = 0$; $\mathrm{p'(k|x_i)}$ is the result of label smoothing; and $\mathrm{w_p}$ denotes the VFRP with a range of (0, 1). The classification loss of VFO is calculated between this assigned ground truth and the model output. The accuracy of mAP of the detection model has been improved by 1.5% through this soft allocation strategy, as shown in the experimental results in Section V.

The classification loss of VFO uses a binary cross-entropy, which can be expressed as

$$L_f = -\sum\nolimits_{i,j} I_{obj}^{ij}\left[\widehat{c_f^{ij}}\log\left(c_f^{ij}\right) + \left(1-\widehat{c_f^{ij}}\right)\log\left(1-c_f^{ij}\right)\right], \tag{3}$$

where I_{obj}^{ij} denotes if the object appears in cell i, j.

b: VFRP

The VFRP is defined as the proportion between the pixel width of the vehicle front or rear and the object's 2D bounding box width, with a data range of (0–1). The purpose of regressing this value is to calculate the vehicle's physical width.

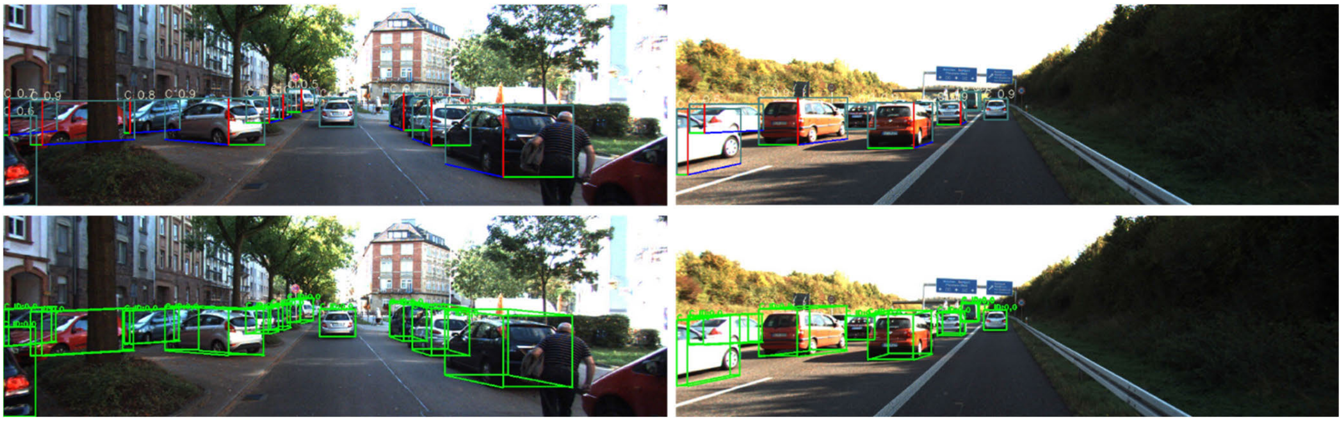

FIGURE 11. Qualitative examples of our proposed method for 3D CAR detection in KITTI-P3D dataset. The top two rows are the pseudo 3D object detection results and the bottom two rows are the corresponding 3D bounding boxes results.

The VFRP regression loss uses MSE, which can be expressed as follows:

$$L_{r_w} = \sum_{ij} I_{obj}^{ij} \left[(r_w^{ij} - \widehat{r_w^{ij}})^2 \right], \tag{4}$$

where $\widehat{r_w^{ij}}$ is the proportion of ground truth, and r_w^{ij} is the model output.

c: VTO

The VTO is defined as the traveling orientation of the vehicle in the image perspective, which corresponds to VFO. The categories of VFO has three properties: left, middle, and right. During training, the VTO label can also exhibit a fuzzy pose, as shown in Fig. 9. Therefore, the label smoothing allocation strategy for VTO is defined as follows:

$$p'(k|x_i) = (1-\alpha)\cdot p(k|x_i) + \alpha/3, \tag{5}$$

$$\alpha = \begin{cases} 0 & if w_p < 0.9 \\ max((w_p - 0.9) * 10, 0) & otherwise, \end{cases} \tag{6}$$

The classification loss of vehicle travel orientation uses the multiclass cross-entropy, which can be expressed as:

$$L_{ori} = -\sum_{i,j} I_{obj}^{ij} \sum_{o\in[l,m,r]} \left[\widehat{c_o^{ij}} \log\left(c_o^{ij}\right) \right], \tag{7}$$

d: VTOP

VTOP is used to determine the global angle by calculating the proportion between the pixel height of the orientation and the object's 2D bounding box height in the image coordinate. The VTOP regression loss uses MSE, which can be expressed as follows:

$$L_{r_h} = \sum_{ij} I_{obj}^{ij} \left[(r_h^{ij} - \widehat{r_h^{ij}})^2 \right], \tag{8}$$

where $\widehat{r_h^{ij}}$ is the ground truth of the VTOP, and r_h^{ij} is the model output.

Finally, the loss function of the VKP-P3D model can be expressed as

$$Loss = \omega_1 \cdot L_{2d} + \omega_2 \cdot \left(L_{dir} + L_{r_w} + L_{ori} + L_{r_h} \right). \tag{9}$$

The loss function utilized for 2D bounding box detection is analogous to the one employed in YOLOv7 [7]. In the context of pseudo 3D object detection, this study reformulates the task of detecting visible key points as a classification problem for VFO and VTO, coupled with regression for VFRP and VTOP. Specifically, the cross-entropy loss function is employed for VFO and VTO classification, whereas the mean squared error loss is applied for VFRP and VTOP regression. The VKP-P3D model is trained by integrating the 2D bounding box loss with the pseudo 3D key point loss, where ω_1 and ω_2 serve as hyperparameters to adjust the loss balance. In this study, ω_1 is set to 1, and ω_2 is assigned a value of 0.6.

3) 3D POSITION MAPPING BASED ON CAMERA GEOMETRY

After regressing the 2D bounding box and pseudo 3D information using the VKP-P3D network, the four key points of the object, p1, p2, p3, and p4, can be obtained, as shown in Fig. 8. These key points belong to the vertices of the object's 3D bounding box, and the three edges correspond to the length, width, and height.

To find the other four vertices of the 3D box, the coordinates of the other vertices in the camera coordinate should be calculated based on cube constraints and the geometric parameters of the monocular camera. The calculation formula is as follows:

$$\begin{bmatrix} x\cdot z \\ y\cdot z \\ z \end{bmatrix}_p = M \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{3d}, \tag{10}$$

where $M \in R^{3\times4}$ is the projection matrix from the 3D location of the camera coordinates into the image. The pixel coordinates of the other four vertices in the image coordinate system are obtained by inverse transformation.

IV. DATASET

Currently, the intelligent driving field primarily utilizes 2D and 3D bounding box detection datasets in image and camera coordinates, respectively. However, 2D object detection datasets only provide bounding box information in

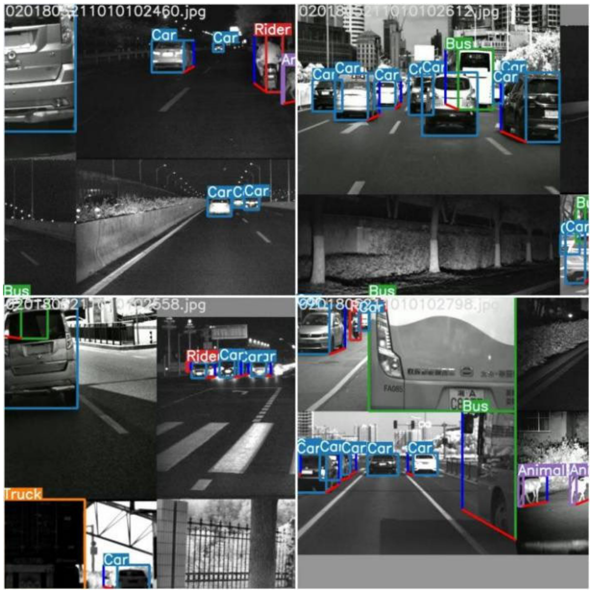

FIGURE 12. Data augmentation while training process. Different from the prior monocular 3D object detection works, the visible key points of proposed model learned are in the image coordinate system. Effective online data augmentation methods validated in 2D object detection model can be used during the training process.

image, which is insufficient for intelligent driving application scenarios. In addition, annotating 3D information for an object based on LiDAR point cloud is costly. Particularly, when the recognition scene is extremely complex, there exists a significant demand for training data volume. At the same time, many lower-level intelligent driving assistance system do not have access to LiDAR sensors. Therefore, this paper constructs the StellaIR-P3D dataset, a pure visual pseudo 3D object detection dataset that annotates the visual 3D box vertices of each object in the image coordinate system. See Fig. 8. The annotation process does not require point cloud data, making this dataset more accessible. Moreover, pseudo 3D information are annotated in the KITTI dataset, namely, KITTI-P3D, which verifies the performance of our proposed VKP-P3D algorithm in this paper.

A. VISIBLE KEY POINTS ANNOTATION

The visible key point-based pseudo 3D object detection algorithm proposed in this paper obtains the object's 3D information by learning the 2D bounding box and the visible 3D box vertices in the image coordinate system. It assumes that the object's 3D box is enclosed by the object's 2D bounding box, and the visible key points in the image coordinate system lie on the edges of the 2D box with geometric constraints.

Therefore, the bounding box annotation software is initially used to annotate the 2D box and categories of the objects. Then, the point annotation software is used to annotate the three visible 3D box vertices of each object on the four boundaries of the object's rectangle, as shown in Fig. 8. These annotated key points are defined as the vertices of the object's 3D box in the image coordinate system, and the VFO label is also tagged.

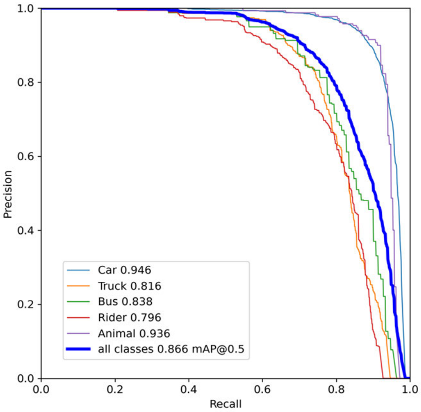

FIGURE 13. P-R recognition curves of VKP-P3D for different categories in StellaIR-P3D dataset.

After annotating the 2D bounding box and visible key points of the pseudo 3D object, ground truth labels must be generated for the objects. In this paper, the label of each object includes the category, the coordinates of the 2D box, c_f of VFO, r_w of VFRP, c_o of VTO, and r_h of VTP.

B. STELLAIR-P3D DATASET

Current publicly available datasets are mainly constructed using visible light cameras, with environmental light conditions and image quality being relatively good. However, the recognition performance under adverse conditions, such as low light, rain, and snow, need to be verified.

TABLE 1. Ablations studies. We ablate the effects of different optimization strategies for VKP-P3D based model.

Model(Small)	mAP(0.5)	mAP(0.5~0.95)
VKP-P3D-ReLU	74.9%	49.3%
VKP-P3D-ReLU-DA	81.6%	51.5%
VKP-P3D-ReLU-DA-Att-ELAN	83.3%	55.3%
VKP-P3D-Att-ELAN-DA-SiLU	84.5%	55.9%
VKP-P3D-Att-ELAN-DA-SiLU-OSA	**84.6%**	**56.5%**

In this paper, a dataset is constructed based on a visual sensor with an active near-infrared light source. The visual system emits a near-infrared light of 808-nm band through a laser module, which is reflected by the objects, perceived by the CMOS sensor and processed by the ISP algorithm to achieve high-definition imaging. Due to the active near-infrared light source, the imaging effect of the IR visual sensor is unaffected by the external environmental light (low light, glare, shadow, rain, and snow). The CMOS sensor only receives the light emitted by the active light source from itself, which can achieve all-weather visual perception in various environments. Example images are shown in Fig. 10.

The active near-infrared visual sensor is installed on the data collection vehicle. The camera parameters are calibrated using a monocular vision calibration method. The dataset is collected under various conditions, including daytime, nighttime, rain, and snow, resulting in a total of 6,000 images annotated according to Step A, which includes object category, 2D bounding box, and visible 3D box vertices. The dataset comprises five categories: Car, Bus, Truck, Rider, and Animal. The dataset is divided into training set, validation set, and test set, with a ratio of 8:1:1.

C. KITTI-P3D DATASET

The KITTI dataset [12] has been widely known in the field of intelligent driving object perception. One of its major strengths lies in its ability to provide comprehensive information for various sensor-based perception algorithms, including one or more cameras and LiDAR sensors. As such, this dataset has become the primary benchmark for evaluating and comparing multiple algorithms, yielding highly promising experimental results. However, the 3D information of the objects in the KITTI monocular 3D object detection dataset is obtained through LiDAR sensors, which is different from the pure visual feature-based 3D object detection task studied in this paper. To further validate the proposed algorithms, we annotate the visible 3D box vertices for the Car category based on the 2D bounding box in the KITTI dataset to construct the KITTI-P3D dataset.

The KITTI-P3D dataset contains 7,481 training images and 7,518 test images. During algorithm validation, given that no test ground-truth labels are available, we divide the training dataset into training and validation sets, following the splitting method in [28]. To ensure robust verification results, we apply the official difficulty settings of easy, moderate, and hard, as described in [12], for accuracy comparisons.

V. EXPERIMENT

A. EXPERIMENTAL SETUP

1) COMPARISON METHOD

To verify the performance of the VKP-P3D algorithm proposed in this paper, we compare its accuracy with YOLOv5 and YOLOv7 object detection algorithms, which are known for their speed and accuracy in current object detection algorithms. However, these algorithms are 2D bounding box detector and cannot output the pseudo 3D information of objects. Therefore, we add the visible key points detection module to the original 2D bounding box detector, namely, YOLOv5-P3D and YOLOv7-P3D. Both are implemented using the PyTorch framework, similar to VKP-P3D. The same hardware platform, NVIDIA 3070 GPU and an Intel I9 CPU, was used for all experiments.

2) HYPERPARAMETER SETTING

The hyperparameter setting during the training process of the proposed VKP-P3D includes 300 epochs with a batch size of 4. A polynomial decay learning rate scheduling strategy is used, with an initial learning rate of 0.01 and a decay rate of 0.1. A warm-up period for 3 epochs is also established. The loss gain for the network regression of the category, object confidence, 2D box, and pseudo 3D attributes is set to 0.3, 0.7, 0.05, 0.02.

3) DATA ONLINE AUGMENTATION

Image augmentation techniques can effectively improve model stability and accuracy without increasing the computational complexity of the model. Different from other vision-based monocular 3D object detection algorithms, the object's ground truth is located in the 3D space of the camera coordinate system. All regression parameters of the VKP-P3D are in the image coordinate. Effective image augmentation methods used in the 2D object detection, including translation, flip, scaling, color space transformation, multiple image mosaic, and image fusion, can be applied to the training process. Image augmentation samples are shown in Fig. 12. The experimental results show that the detection accuracy of the VKP-P3D algorithm is improved by 6.7% mAP through these image augmentation methods, as demonstrated in TABLE 1.

B. EXPERIMENTAL RESULT

To verify the detection ability of the proposed VKP-P3D, the performance of the various aforementioned optimization strategies is initially verified through ablation experiments, and the optimal model structure is established.

On the basis of this optimal structure, different depth and width networks, such as VKP-P3D-S (small, 6.12 GMac FLOPs), VKP-P3D-M (moderate, 13.73 GMac FLOPs), and VKP-P3D-L (large, 21.24 GMac FLOPs), are constructed. Then, the performance is compared with YOLOv5-P3D and YOLOv7-P3D algorithms on StellaIR-P3D and KITTI-P3D datasets, respectively. The evaluation criterion of this paper is the mAP of the 2D object bounding box via intersection over union (IoU) and the object keypoint similarity (OKS) of visible 3D box key points [14]. The experimental results are described as follows.

1) COMPARISON OF ABLATION EXPERIMENTAL RESULTS

This study conducts ablation experiments on the StellaIR-P3D dataset using the proposed VKP-P3D-S model. It compares the effects of various optimization strategies, using mAP as the key metric.

VKP-P3D-ReLU: No augmentation data are used during the training process with ReLU activation function.

VKP-P3D-ReLU-DA: During the training process, data augmentation operations mentioned in the V-A are performed online with random parameters.

VKP-P3D-ReLU-DA-Att-ELAN: The Att-ELAN structure is added on the previous model.

VKP-P3D-Att-ELAN-DA-SiLU: ReLU activation function is replaced by the Sigmoid linear unit (SiLU) [49].

VKP-P3D-Att-ELAN-DA-SiLU-OSA: The OSA structure is added on the previous model.

The results of the ablation experiments are shown in TABLE 1.

2) 2D BOUNDING BOX IOU

The 2D bounding box IoU is an important criterion for evaluating the performance of visual object bounding box

TABLE 2. Comparison 2D object detection mAP in StellaIR-P3D dataset.

Model		#Param.	FLOPs	Size	FPS_{3070}	mAP_{50}	$\text{mAP}_{50:95}$
YOLOv5-P3D [6]	YOLOv5-P3D-S	4.83M	5.92GMac	640	93	75.7%	47.0%
	YOLOv5-P3D-M	11.43M	13.33GMac	640	82	77.4%	50.4%
	YOLOv5-P3D-L	17.15M	21.21GMac	640	74	79.4%	52.9%
YOLOv7-P3D [7]	YOLOv7-P3D-S	3.66M	5.93GMac	640	103	82.8%	55.3%
	YOLOv7-P3D-M	9.34M	13.29GMac	640	100	85.4%	59.1%
	YOLOv7-P3D-L	14.1M	20.59GMac	640	99	85.9%	59.0%
VKP-P3D(Ours)	VKP-P3D-S	4.08M	6.12GMac	640	102	84.6%	56.4%
	VKP-P3D-M	10.43M	13.73GMac	640	99	86.5%	59.8%
	VKP-P3D-L	15.73M	21.24GMac	640	97	**86.6%**	**60.4%**

TABLE 3. Comparison AP(Average Precision) of each category in StellaIR-P3D dataset.

Model		Car	Truck	Bus	Rider	Animal
YOLOv5-P3D [6]	YOLOv5-P3D-S	91.0%	69.4%	65.5%	69.0%	83.6%
	YOLOv5-P3D-M	91.0%	71.3%	66.0%	70.7%	88.2%
	YOLOv5-P3D-L	91.4%	73.2%	70.0%	73.4%	89.2%
YOLOv7-P3D [7]	YOLOv7-P3D-S	93.7%	78.7%	76.8%	76.0%	90.0%
	YOLOv7-P3D-M	94.4%	81.8%	80.4%	77.4%	**94.6%**
	YOLOv7-P3D-L	**94.7%**	81.2%	82.1%	79.3%	93.7%
VKP-P3D(Ours)	VKP-P3D-S	93.9%	80.9%	79.6%	76.0%	92.7%
	VKP-P3D-M	**94.7%**	**82.5%**	82.6%	79.2%	93.8%
	VKP-P3D-L	94.6%	81.6%	**83.8%**	**79.6%**	93.6%

TABLE 4. Comparison AP in car category of KITTI-P3D dataset with different difficulty set.

Model		FPS_{3070}	Easy (AP50)	Moderate (AP50)	Hard (AP50)
YOLOv5-P3D [6]	YOLOv5-P3D-S	121	84.6%	82.4%	82.2%
	YOLOv5-P3D-M	107	86.1%	84.9%	83.1%
	YOLOv5-P3D-L	96	86.0%	85.4%	83.8%
YOLOv7-P3D [7]	YOLOv7-P3D-S	134	84.1%	81.3%	82.2%
	YOLOv7-P3D-M	130	86.5%	84.6%	83.9%
	YOLOv7-P3D-L	129	87.0%	85.1%	84.2%
M3D-RPN [28]	-	14	**90.2%**	83.7%	67.7%
VKP-P3D(Ours)	VKP-P3D-S	133	86.5%	83.3%	82.8%
	VKP-P3D-M	129	88.1%	**85.2%**	84.8%
	VKP-P3D-L	126	88.6%	84.9%	**85.0%**

detection algorithms. It determines whether the object is detected by the IoU threshold between the 2D box detection result and the ground truth box. Then, it calculates the mAP by precision and recall (P–R) curve.

In the StellaIR-P3D dataset, the mAP of the proposed VKP-P3D is compared with those of YOLOv5-P3D [6] and YOLOv7-P3D [7], as shown in TABLE 2. The average detection precision for each category of the compared algorithms is demonstrated in TABLE 3. The P–R curve of the VKP-P3D-L model is shown in Fig. 13.

In the KITTI-P3D dataset, the average precision (AP) of the proposed method and compared models in detecting the Car category is shown in TABLE 4 under different difficulty levels (easy, moderate, and hard).

TABLE 5. Comparison the visible key points detection AP with OKS>0.5 in StellaIR-P3D dataset.

Model	AP50 (OKS>0.5)
3D-like [34]	71.6%
YOLOV5-S [6]	79.2%
YOLOV7-S [7]	84.7%
VKP-P3D-S (Ours)	**85.1%**

From the results of the comparative experiments on the two datasets, the proposed VKP-P3D algorithm has higher detection precision under a similar detection speed.

3) VISIBLE KEY POINTS SIMILARITY

The object visible key points coordinates obtained by the proposed VKP-P3D are all in image coordinate system and cannot use the mAP of 3D IoU as the accuracy metric. In the 2D pose estimation challenge, the COCO evaluation defines OKS as the main competition metric for localizing anatomical key points [14]. OKS plays the same role as the IoU in object detection. It is calculated from the scale of the object and the distance between the predicted and ground truth points. Therefore, for evaluating the accuracy of pseudo 3D object detection, the OKS criterion of visible key points is used as the measurement metric, with the calculation formula as follows:

$$OKS = \frac{\sum_{i} \left[exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0) \right]}{\sum_{i} [\delta(v_i > 0)]}, \qquad (11)$$

where d_i is the Euclidean distance of each key point between the corresponding ground truth and detector output. v_i is the point visibility flag of the ground truth. In this paper, all key points are visible. Thus, v_i is consistently equal to 1. s is the area of the object bounding box. k_i is a constant that controls the falloff. We set k_i to 0.067 in this paper. The AP for the visible key points detection of the proposed VKP-P3D in the StellaIR-P3D dataset is shown in TABLE 5. Experiments showed that the proposed detector has a significant increase in the accuracy of key points detection.

Figs. 10 and 11 demonstrate the detection results of the VKP-P3D algorithm in the StellaIR-P3D and KITTI-P3D datasets, where the above two rows represent the pseudo 3D detection results, and the bottom two rows represent the detection results of converting the object's pseudo 3D information to the 3D object in the camera coordinate system.

VI. CONCLUSION

In this paper, we propose a visible key points based pseudo 3D object detection algorithm, namely, VKP-P3D, for intelligent driving scenarios. This algorithm converts the 3D object detection problem into 3D box visible key points detection within the image coordinate system. It constructs a real-time monocular 3D object detection model through a single-shot RPN network. In comparison with 2D object detection, the VKP-P3D provides richer object information for the decision module of the intelligent driving system. In comparison with previous works of monocular 3D object detection, the proposed VKP-P3D does not require auxiliary networks or multilevel structures. In addition, during the annotation process of the training dataset based on this algorithm, no LiDAR point cloud information is required, thereby greatly reducing the cost of data acquisition.

In terms of feature extraction, we propose an Att-ELAN structure, which enhances key granularity feature channels with attention structures for different channels. It also weakens invalid features and improves the model's expressive ability in complex traffic scenes.

To verify the algorithm's accuracy, we construct the StellaIR-P3D dataset, featuring annotated object's 2D bounding box and visible 3D box key points, with categories of Car, Bus, Truck, Rider, and Animal. At the same time, we construct the KITTI-P3D dataset with the Car category of the official KITTI dataset. Then, we conduct experiments and verification on the two datasets with the proposed VKP-P3D algorithm and compared it with YOLOv5-P3D and YOLOv7-P3D. The experimental results show that the proposed VKP-P3D has excellent detection accuracy and speed.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[2] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 920–933, Apr. 2017.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[4] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[5] Y. Liao, G. Chen, and R. Xu, "Enhanced sparse detection for end-to-end object detection," *IEEE Access*, vol. 10, pp. 85630–85640, 2022.

[6] [Online]. Available: https://github.com/ultralytics/yolov5

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[8] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.

[9] W. Wang and Y. Gou, "An anchor-free lightweight object detection network," *IEEE Access*, vol. 11, pp. 110361–110374, 2023.

[10] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[11] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617011.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[13] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.

[14] *MSCOCO Keypoint Evaluation Metric*. [Online]. Available: http://mscoco.org/dataset/#keypoints-eval

[15] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[16] H. Hu, M. Zhu, M. Li, and K.-L. Chan, "Deep learning-based monocular 3D object detection with refinement of depth information," *Sensors*, vol. 22, no. 7, p. 2576, Mar. 2022.

[17] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[18] Y. Wu, Y. Wang, S. Zhang, and H. Ogai, "Deep 3D object detection networks using LiDAR data: A review," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1152–1171, Jan. 2021.

[19] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 5291–5304, 2022.

[20] C. Chen, L. Z. Fragonara, and A. Tsourdos, "RoIFusion: 3D object detection from LiDAR and vision," *IEEE Access*, vol. 9, pp. 51710–51721, 2021.

[21] N. A. M. Mai, P. Duthon, L. Khoudour, A. Crouzil, and S. A. Velastin, "3D object detection with SLS-fusion network in foggy weather conditions," *Sensors*, vol. 21, no. 20, p. 6711, Oct. 2021.

[22] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21653–21662.

[23] H. A. Hoang and M. Yoo, "3ONet: 3-D detector for occluded object under obstructed conditions," *IEEE Sensors J.*, vol. 23, no. 16, pp. 18879–18892, Aug. 2023.

[24] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6177–6186.

[25] Y. Chen, S. Liu, X. Shen, and J. Jia, "DSGN: Deep stereo geometry network for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12533–12542.

[26] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5632–5640.

[27] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6850–6859.

[28] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9286–9295.

[29] H. Hu, M. Li, M. Zhu, W. Gao, P. Liu, and K.-L. Chan, "Monocular 3D object detection with motion feature distillation," *IEEE Access*, vol. 11, pp. 82933–82945, 2023.

[30] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, "GS3D: An efficient 3D object detection framework for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1019–1028.

[31] L. Liu, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1057–1066.

[32] F. Manhardt, W. Kehl, and A. Gaidon, "ROI-10D: Monocular lifting of 2D detection to 6D pose and metric shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2064–2073.

[33] Y. Liu, Y. Yixuan, and M. Liu, "Ground-aware monocular 3D object detection for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 919–926, Apr. 2021.

[34] C. Wang, L. Zhou, J. Li, and W. Yang, "3D-like bounding box for vehicle detection," in *Proc. 34rd Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Jun. 2019, pp. 244–249.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8691. Springer, 2014.

[36] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7029–7038.

[37] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10208–10219.

[38] Anonymous. *Designing Network Design Strategies*, 2022.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[40] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 752–760.

[41] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[43] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.

[44] S. Li, C. He, R. Li, and L. Zhang, "A dual weighting label assignment scheme for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9377–9386.

[45] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 303–312.

[46] W. Xiao, H. Liu, Z. Ma, W. Chen, C. Sun, and B. Shi, "Fatigue driving recognition method based on multi-scale facial landmark detector," *Electronics*, vol. 11, no. 24, p. 4103, Dec. 2022.

[47] S. Du, H. Wang, Z. Yuan, and T. Ikenaga, "Bi-pose: Bidirectional 2D-3D transformation for human pose estimation from a monocular camera," *IEEE Trans. Autom. Sci. Eng.*, early access, Jun. 1, 2023, doi: 10.1109/TASE.2023.3279928.

[48] W. Bao, Z. Ma, D. Liang, X. Yang, and T. Niu, "Pose ResNet: 3D human pose estimation based on self-supervision," *Sensors*, vol. 23, no. 6, p. 3057, Mar. 2023.

[49] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.

CHANGLIANG SUN received the M.S. degree from the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China, in 2016. He is currently pursuing the Ph.D. degree in electronic science and technology with the College of Electrical and Information Engineering, Hunan University, Changsha, China.

His research interests include intelligent driving perception, machine vision, and deep learning.

HONGLI LIU (Member, IEEE) received the B.Sc. degree in electrical engineering and the Ph.D. degree in control theory and engineering from Hunan University, Changsha, China, in 1985 and 2000, respectively.

He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include machine vision, intelligent information processing, and transmission technology.

WEICHU XIAO received the B.S. degree in electronic information engineering from Hunan Normal University, Changsha, China, in 1998, and the M.S. degree in electronics and communication engineering from Hunan University, Changsha, in 2012, where he is currently pursuing the Ph.D. degree in energy and power with the College of Electrical and Information Engineering.

His research interests include intelligent information processing, driver-distracted detection, and cyber-physical systems.

BO SHI received the M.S. degree in electronic science and technology from Hunan University, Changsha, China, in 2020, where he is currently pursuing the Ph.D. degree in electronic science and technology with the College of Electrical and Information Engineering.

His research interests include machine vision, optical three-dimensional measurement, and railway inspection.

YUAN QIU is currently pursuing the Ph.D. degree with Hunan University, Changsha, China.

She is with the College of Electrical and Information Engineering, Hunan University. Her research interests include machine vision, image processing, computer vision, machine learning, and railway fastener inspection.